Evaluating the energy impact of device parameters for DNN inference on edge

Anurag Dutt*, Sri Pramodh Rachuri*, Ashley Lobo, Nazeer Shaik, Anshul Gandhi, Zhenhua Liu
Stony Brook University, Stony Brook, NY, 11794
{anurag.dutt, sripramodh.rachuri, ashley.lobo, nazeer.shaik, anshul.gandhi, zhenhua.liu}@stonybrook.edu

I. Introduction

In recent years, Deep Neural Networks (DNN), including Large Language Models (LLMs), have gained significant traction. While the training of these models has received attention, research on efficient deployment for *inference*, especially at the *edge*, is still ongoing [2], [9]. The edge devices are often deployed using batteries or solar panels for various critical applications in remote locations, which motivates our investigation in this paper on the *energy consumption* for these devices [8]. Efficient deployment of DNN on edge faces challenges due to the complex models straining the limited edge resources and the need to tune the various available hardware parameters. In fact, parameters like GPU frequency can impact energy, power, and inference time. This impact could also be *non-monotonic* making it difficult to find the optimal settings. Consequently, practitioners often settle for sub-optimal energy savings by not carefully tuning these parameters and resorting to existing tools, such as Dynamic Voltage and Frequency Scaling (DVFS), for configuring hardware knobs [16].

In this paper, we study the impact of hardware knobs on the energy consumption of DNN inference, specifically for edge devices. There has been some related work recently that looks at the energy efficiency of DNN inference. Holly *et al.* [4] profile Mobilenet-V2 as a function of hardware parameters (*e.g.*, number of cores); however, the workloads are limited to CNNs. DeepEdgeBench [1] analyzes the power consumption of running DNN models on different edge devices, but the authors do not investigate the impact of hardware parameter changes on energy. Likewise, there has been a lot of recent work on analyzing energy-efficient *training* of DNN workloads on edge (*e.g.*, Prashanthi *et al.* [12], Trainer [15], EfficientGrad [5]) and on servers (*e.g.*, Zeus [16]). However, insights from energy analysis for training need not translate to inference since training is more sensitive to memory, network delays, and parallelization.

We conduct our empirical study using smart edge devices with compact ARM-based microprocessors with GPU acceleration. For our experiments, we use two devices: (1) Jetson Nano, an entry-level edge device, and (2) Jetson Xavier NX, a mid-tier, more powerful version. Both devices are capable of serving DNN models for a variety of practical application tasks. Our experimental results show that we can reduce inference energy by as much as 19% compared to DVFS by optimally tuning the CPU and GPU frequencies.

This work was supported by NSF grants 2214980, 2106434, and 1750109.
*First two authors contributed equally to this work.

II. Experiment Design

Both Jetson Nano and Xavier NX provide multiple onboard sensors measuring power for different components and can be accessed through an `I2C` interface. We log power consumed by the entire module every 100ms. The energy overhead of 100ms polling was less than 0.5%. We found that more frequent polling (*e.g.*, 10ms or 1ms) can result in higher energy overheads of more than 2%.

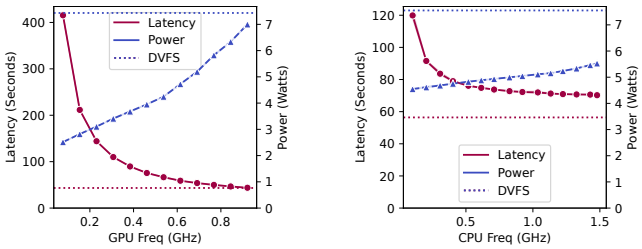
Both devices have three main hardware knobs that can be changed during runtime—CPU clock frequency, GPU clock frequency, and the number of cores. We focus on the first two since the number of cores did not impact the power consumption in our experiments as the DNN models are implemented in Python, which is effectively single-threaded due its Global Interpreter Lock (GIL) mechanism. Both devices offer a large range of CPU and GPU frequencies. Jetson Nano has 180 possible GPU and CPU frequency combinations and Xavier NX has 375 possible configurations. We use `sysfs` APIs to set static CPU and GPU frequencies. Baselines were established using the default DVFS module.

We executed the DNN inference process and logged the workload execution checkpoints and their requisite timestamps while recording the power metrics on a separate thread. Merging these logs and correlating them with inference events (*e.g.*, model initialization, inference start) allowed us to compute the power, energy, and execution time for the inference workload. During testing, we disconnected all peripherals from the device and killed unnecessary background tasks.

Each experiment consisted of running a DNN inference workload under a specific CPU and GPU frequency setting. For our workloads, we used AlexNet [7], ResNet-18 [3], and MobileNet-v2 [10] for image classification, YOLOv4-Tiny [6] for object detection, and DistilBERT [11] and BERT-Tiny [13] for natural language classification. The workload in every experiment was kept constant at 3,200 inferences. Each experiment was repeated 10 times and average values were reported; the variation between runs was low (less than 5%).

III. Evaluation Results

Frequency Scaling Sweep and DVFS: We start by analyzing the impact of GPU frequency scaling on power and inference time, as shown in Figure 1(a) for an AlexNet inference workload on Jetson Nano for a batch size of 16. We see



(a) Changing GPU frequency (b) Changing CPU frequency

Fig. 1: Comparison of inference latency and power consumed under different CPU/GPU frequencies when running AlexNet on Jetson Nano.

that power consumption increases almost linearly with GPU frequency. However, the *decrease in inference latency starts to plateau out at higher GPU frequencies* as GPU is no longer the bottleneck and the performance is limited by other components such as memory and I/O bandwidth; similar effects have been noted in prior work on servers [14]. Both power and latency vary greatly as GPU frequency changes from lowest to highest (CPU frequency was fixed at 1132.8 MHz); the span of power consumed is 4.7W, whereas the span of latency is 532s.

Figure 1(b) shows a similar result, but for changing CPU frequencies; here, we fix the GPU frequency at 0.6912 GHz. While the trend is somewhat similar, we see that the span of power (1W) and span of latency (60s) is much more narrower, indicating that ***CPU frequency has a smaller impact compared to GPU frequency***. This is because computationally intensive operations, such as tensor operations, are done by the GPU, whereas the CPU is responsible for less intensive tasks such as data preprocessing, initialization, and control flow.

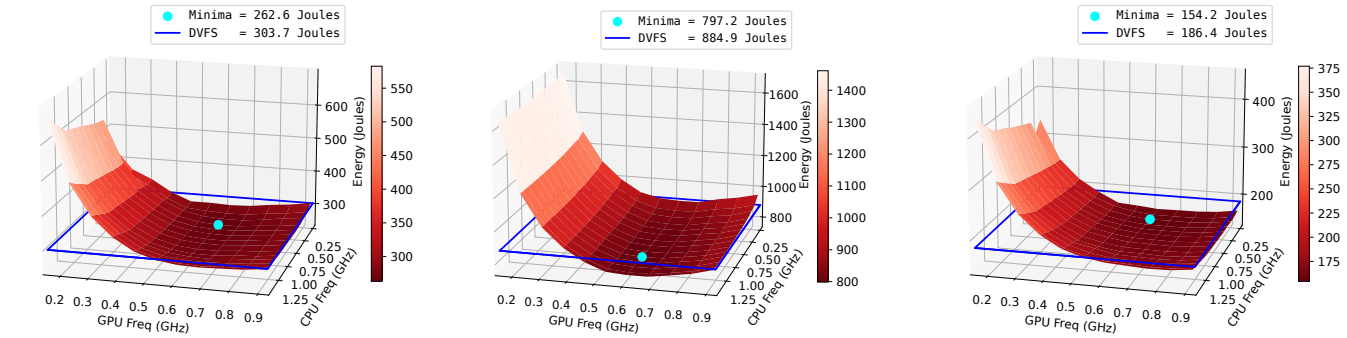
Next, we evaluate the impact of DVFS on power and latency. The dashed lines in Figure 1 show the behavior when using DVFS (defaults are 'nvhost_podgov' for GPU and 'schedutil' for CPU) instead of manually setting frequencies. We see that DVFS affords low latency but incurs very high power consumption, suggesting that *default DVFS opts for higher frequencies*. This was confirmed by profiling, which revealed that DVFS operated at the highest CPU and GPU frequencies 89% and 83% of the time, respectively.

We also experimented with the other DVFS governors available (*e.g.*, 'powersave', 'performance', 'ondemand'). We found that 'powersave' consumes <0.01% more energy as compared to the default governors, primarily because although the average power is reduced by 30%, the execution time of the workload is increased by 43%. Likewise, we found that all other DVFS governors perform similarly to the default governor, with the difference in energy consumption being within 1%. With CPU DVFS, the best among other governors had 2.9% lower power but with 2.6% higher inference latency.

Energy Topology under Jetson Nano: To study the impact of energy, we plot the energy consumed for inference as a function of CPU and GPU frequency for 3 DNN workloads for Jetson Nano in Figure 2. The batch size is fixed at 16 for all workloads. We see that energy does not change much with CPU frequency for a given GPU frequency. However, for a given CPU frequency, the ***energy consumption does change substantially with GPU frequency***. While not always visible, the *impact of GPU frequency on energy is not monotonic*; there exist some moderately high GPU frequencies at which the energy is minimized, as shown by the cyan dot in the plots.

The minimum energy obtained by sweeping over all the CPU and GPU frequencies shows that energy can be lowered significantly when compared to DVFS, as noted in the legend for each subfigure. The percentage reductions in energy afforded by the minima over DVFS for Figures 2(a)–2(c) are 13.5%, 9.9%, and 17.2%, respectively. We also find that the percentage reductions afforded by the minima over DVFS for our other three models are 19.3% for ResNet-18, 15.6% for MobileNet-v2, and 17.3% for distilBERT. However, we found that the latency under the minima configuration is typically 28%–35% higher than that achieved under DVFS. The minima usually occurred at the GPU frequency of 614.4 MHz, with the optimal CPU frequency varying across the workloads.

Energy Topology under Jetson Xavier NX: We ran similar experiments for Xavier NX for all the workloads. We found similar trends as with the Nano (thus omitting figures), with the minima affording an energy reduction in the range of 2.5%–15.1% across workloads. The energy minima usually occured at a GPU frequency of 803.25 MHz. We also found that, for a given workload and frequency settings, the ***energy consumption of Xavier NX is significantly lower than Nano***, often by at least 2×, and sometimes as much 4×. Given the newer generation of Xavier NX, this is not wholly unexpected.



(a) AlexNet (b) YOLOv4 - Tiny (c) BERT-Tiny

Fig. 2: Inference energy consumption as a function of CPU and GPU frequency under Jetson Nano; the color shade denotes the energy consumption, also shown on z-axis. Also highlighted is the point at which energy is minimized. The blue line represents the energy consumed under DVFS.

REFERENCES

[1] S. Baller, A. Jindal, M. Chadha, and M. Gerndt. Deepedgebench: Benchmarking deep neural networks on edge devices. In *2021 IEEE International Conference on Cloud Engineering (IC2E)*, pages 20–30, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society.

[2] Linux Foundation. Sharpening the edge: Overview of the lf edge taxonomy and framework, Jul 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Stephan Holly, Alexander Wendt, and Martin Lechner. Profiling Energy Consumption of Deep Neural Networks on NVIDIA Jetson Nano. In *Proceedings of the 11th International Green and Sustainable Computing Workshops (IGSC)*, pages 1–6, 2020.

[5] Ziyang Hong and C. Patrick Yue. Efficient-grad: Efficient training deep convolutional neural networks on edge devices with gradient optimizations. 21(2), feb 2022.

[6] Zicong Jiang, Liquan Zhao, Shuaiyang Li, and Yanfei Jia. Real-time object detection method based on improved yolov4-tiny, 2020.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[8] Yu-Jen Ku and Sujit Dey. Sustainable vehicular edge computing using local and solar-powered roadside unit resources. In *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, pages 1–7, 2019.

[9] Sri Pramodh Rachuri, Francesco Bronzino, and Shubham Jain. Decentralized modular architecture for live video analytics at the edge. In *Proceedings of the 3rd ACM Workshop on Hot Topics in Video Analytics and Intelligent Edges*, HotEdgeVideo '21, page 13–18, New York, NY, USA, 2021. Association for Computing Machinery.

[10] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019.

[11] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.

[12] Prashanthi S.K, Sai Anuroop Kesanapalli, and Yogesh Simmhan. Characterizing the performance of accelerated jetson edge devices for training deep learning models. *Proc. ACM Meas. Anal. Comput. Syst.*, 6(3), dec 2022.

[13] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models, 2019.

[14] Qiang Wang and Xiaowen Chu. Gpgpu performance estimation with core and memory frequency scaling. *IEEE Transactions on Parallel and Distributed Systems*, 31(12):2865–2881, 2020.

[15] Yang Wang, Yubin Qin, Dazheng Deng, Jingchuan Wei, Tianbao Chen, Xinhan Lin, Leibo Liu, Shaojun Wei, and Shouyi Yin. Trainer: An energy-efficient edge-device training processor supporting dynamic weight pruning. *IEEE Journal of Solid-State Circuits*, 57(10):3164–3178, 2022.

[16] Jie You, Jae-Won Chung, and Mosharaf Chowdhury. Zeus: Understanding and optimizing GPU energy consumption of DNN training. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, pages 119–139, Boston, MA, April 2023. USENIX Association.